RSAC 2025 Conference

Many Voices.
One Community.

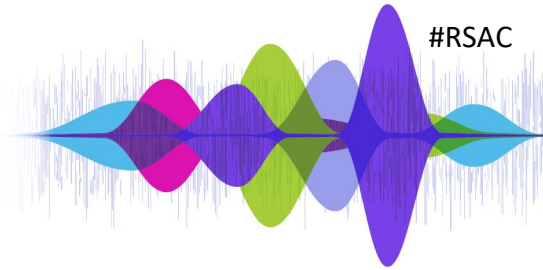
SESSION ID: SAT-W09

Policy-Driven Guardrails for Securing Federated Multi-Modal AI Systems

Magesh Dhanasekaran

Senior Security Consultant
Amazon Web Services
https://www.linkedin.com/in/mageshdh/

Disclaimer



Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference LLC does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

© 2025 RSA Conference LLC or its affiliates. The RSAC and RSAC CONFERENCE logos and other trademarks are proprietary. All rights reserved.

Multi-Modal Federated AI: The New Attack Surface

Core Policy Challenges

Policy Consistency
- Uniform enforcement
- Autonomous interpretations

Modality Interactions
- Cross-modal validation
- Secure communication

Design Considerations:
- Distributed policy enforcement
- Agent boundaries
- Version control
- Trust domains

Framework Requirements

Policy Architecture
- Centralized definition
- Distributed enforcement
- Cross-domain validation

Compliance Framework
- Coherence
- Auditing
- Governance

Security Design
- Zero trust architecture
- Adaptive controls

Federated Multi-Modal Attack: Healthcare Example

Hospital Nodes A, B, ..

- Imaging Data
- Patient Records

Federation Server

- Model Aggregation
- Update Distribution

Attack Surface

- Gradient Poisoning
- Time Desync Attack

Detection Challenges

- Evades IDS
- Multi-Modal Blindspots

Blast Radius

- All Nodes
- Model Corruption
- Trust Breakdown

Incident Response

- Model Quarantine
- Rollback
- Trust Rebuild

VOTING

What's your current level of concern about AI security in your organization?

- High priority - actively addressing it now
- On our radar - planning implementation
- Just beginning to evaluate risks
- Not yet a focus area for us

Policy-Driven Architecture: Beyond Traditional Defenses

Advanced Policy Architecture

1. Adaptive Modality Protection
   Semantic defense layer design
2. Cross-Domain Policy Enforcement
   Modality interaction security
3. Policy Framework
   Designed for adaptability

Design Principles

1. Policy Integration Strategy
   Layered validation approach
2. Cross-Domain Security
   Comprehensive coverage design
3. Performance Considerations
   Balance security and efficiency

Key Architectural Features

Validation Architecture | State-Aware Design | Defense-in-Depth Approach

Policy Implementation: The Devil in the Details

Implementation Challenges

Race Conditions

- Concurrent Updates
- Policy Conflicts

State Management

- Cache
- Cross-model sync

Implementation Targets

Development Goals

- Rapid Rollback
- Reliable Updates
- Version Sync

Implementation Framework

2-phase validation — Zero-downtime deploy — Real-time monitoring

VOTING

How would you describe your organization's approach to AI policy management?

- Comprehensive framework in place
- Partial implementation with some gaps
- Adapting existing security policies for AI
- Still in planning/evaluation phase
- Just starting our AI journey

Adaptive Policy Systems: Evolution in Production

Adaptation Mechanisms

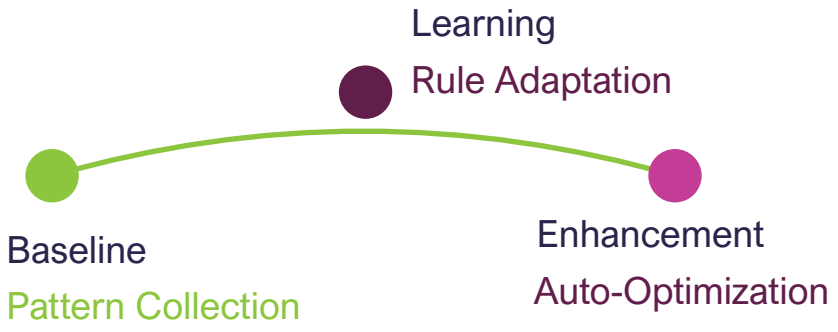
Policy Learning Design

- Pattern Recognition System
- Adaptive Detection Patterns

Defense Perimeter

- Dynamic Boundary System
- High-accuracy Pattern Matching

Evolution Stages



Development Objectives

Response Goals

- Real-time Adaptation
- Continuous Learning Cycles
- High-accuracy Pattern Matching

System Targets

- Minimal False Positives
- Near-perfect Policy Sync
- Rapid Recovery

Enhancement Goals

- Auto-Optimization
- Predictive Defense
- Dynamic Scaling

Real-World Integration: Policy Enforcement in Production

Security Integration Points

Policy Enforcement Points

- Pre-execution validation
- Runtime policy checks
- Post-execution verification

Security Controls

Defense-in-Depth Strategy

- Input sanitization and validation
- Cross-modal security checks
- Policy violation prevention
- Automated threat response

Security Implementation Framework

Prevention Layer

- Policy-driven blocks
- Behavioral analysis
- Proactive protection

Detection Layer

- Real-time monitoring
- Threat pattern matching
- Anomaly detection

Response Layer

- Automated containment
- Incident response
- Forensic logging

VOTING

Last quick poll

Which stage best describes your AI incident response capabilities?

- Mature - AI-specific procedures implemented
- Developing - updating existing IR for AI
- Planning - actively designing AI response
- Evaluating - assessing our needs
- Not started - need guidance on where to begin

Multi-Modal Incident Response: Initial Response Framework

Critical Response Timeline

Initial Response Phase: Immediate Actions

- Policy isolation
- State capture

Assessment Phase: Impact Analysis

- Impact analysis
- Federation state analysis

Containment Phase: Response Actions

- Quarantine
- Safe-mode
- Evidence preservation

Automated Response Matrix

Policy Breach	➡	Immediate Response
Cross-Modal Attack	➡	Rapid Containment
Federation Compromise	➡	Federation-wide Response

Critical Decision Points

Blast Radius Control

- Modality isolation
- Federation partition

Recovery Initialization

- Rollback triggers
- State reconstruction paths

Implementing Multi-Modal Guardrails: Day One Action Plan

Day One Implementation

1. Policy Foundation (2 hours)

- Base policy template deployment
- Initial threshold configuration
- Monitoring instrumentation

2. Agent Integration (4 hours)

- Communication layer setup
- State management initialization
- Basic enforcement activation

3. Validation Checkpoints

- Policy propagation verification
- Basic enforcement testing
- Performance baseline capture

Day One Outcomes

1. Security Posture

- Basic policy enforcement active
- Initial modality protection
- Foundation for evolution

2. Next Steps Preview

- Advanced policy refinement
- Adaptation tuning
- Full federation enablement

The Road Ahead: Immediate Actions & Future Threats

This Week's Critical Tasks

1. Policy Hardening

- Threshold optimization
- Cross-modal rule tightening
- Response time tuning

2. Federation Strengthening

- Agent trust verification
- Communication encryption
- State synchronization

3. Emerging Threat Landscape

- Cross-modal poisoning
- Federation trust exploitation
- Temporal consistency attacks

90-Day Security Roadmap

Month 1: Foundation

- Full policy deployment
- Performance optimization

Month 2-3: Evolution

- Incident response testing
- Advanced threat modeling
- Autonomous adaptation
- Cross-org federation

Development Goals

Policy Management

- Centralized control
- Automated distribution
- Version tracking

Threat Response

- Pattern recognition
- Automated containment
- Evolution tracking

RSAC™ 2025 Conference

Q&A

Many Voices.
One Community.

RSAC™ 2025 Conference

Many Voices.
One Community.

Thank you